# project 1

Piere 2021-11-14

This project has two parts. The first one about apartment prices and the second about covid cases.

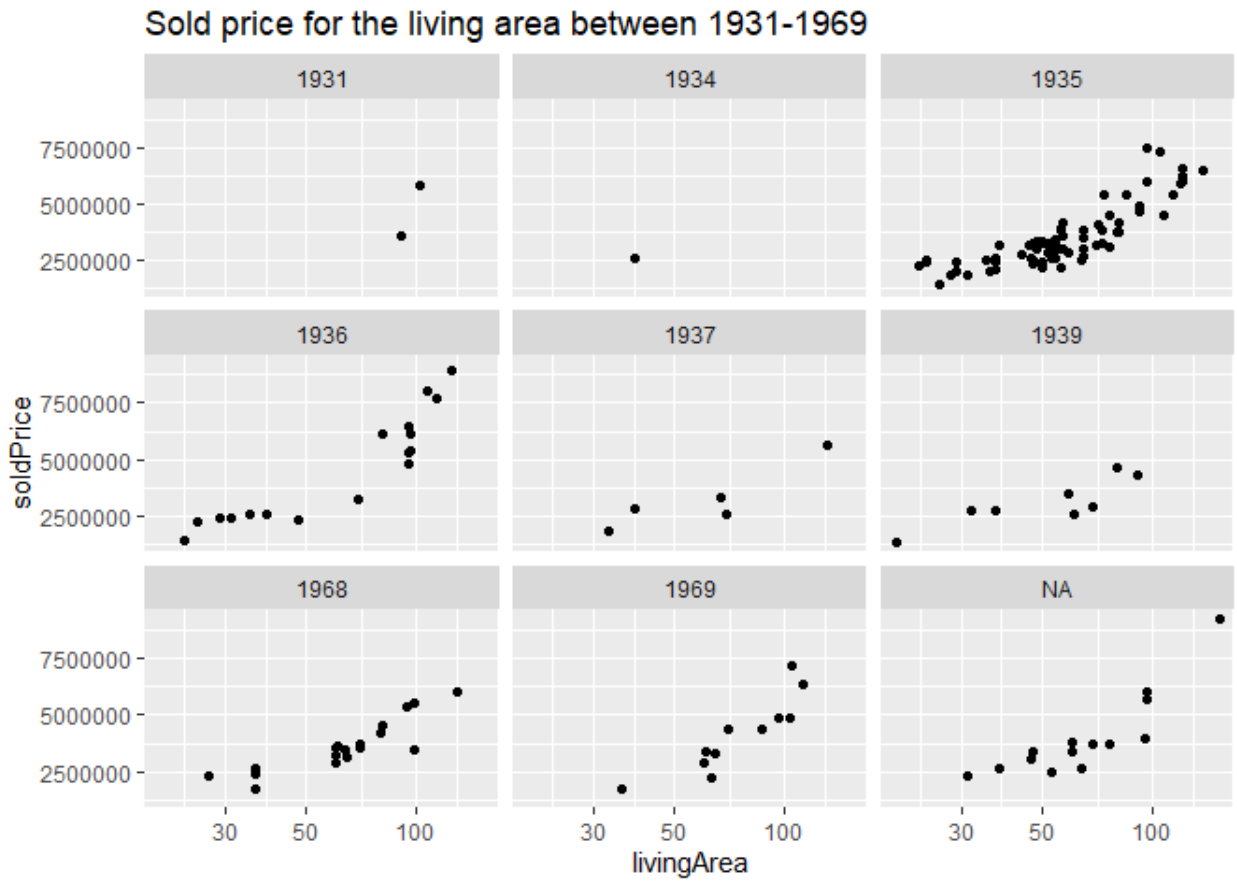
Exercise 1: Apartment prices

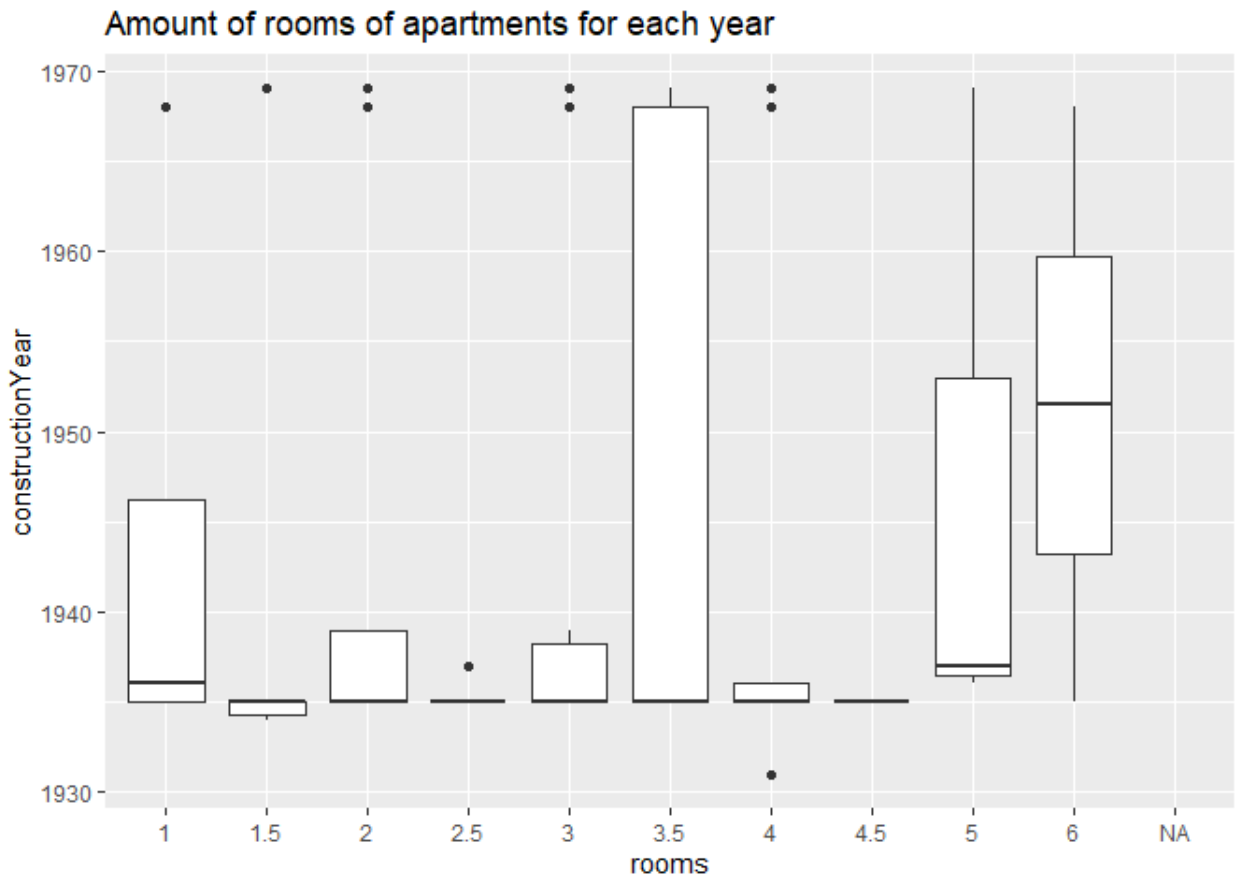
The first exercise contains sales from data on 158 apartments in Ekhagen and we are going to fullfill the following tasks:

Illustrate how Soldprice depends on Livingarea with a suitable figure. Illustrate trends in Soldprice / Livingarea over the period. Illustrate an aspect of data using a table. Illustrate an aspect of data using a boxplot (geom_boxplot).

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 -
## v ggplot2 3.4.0      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() -
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```


Sold price for the living area
soldPrice
7500000
5000000
2500000
30
50
100
livingArea


Sold price for the living area between 1931-1969
1931
1934
1935
1936
1937
1939
1968
1969
NA
soldPrice
7500000
5000000
2500000
30
50
100
livingArea




The first plot shows livingArea against the soldPrice. I decided to go with a scatter plot because it shows the relation between livingArea and soldPrice. We can see that the bigger the area is for the apartment the expensier it becomes. Some points tends to a vertical line which says that only the soldPrice changes and the livingArea is the same. This could be explained by others factors such as year of the building or where the building is located etc. The second plot shows a division of the sold price against the living area but for each construction year. We can see that in year 1935 were the highest amount of apartments sold under 5 million and year 1934 was the lowest. In the third plot i decided to illustrate the data of rooms of apartments for each year as a box plot. We can see that it is very common for an apartment to have 2-3 rooms. The as.factor(contructionYear) in the x axis divides the years so for each year we want to stack the number of rooms there.

Exercise 2: Folkhälsomyndigheten COVID cases and why excel might not be your friend

This second exercise contains data on COVID-19 cases in Sweden. The data was obtained through Folkhälsomyndigheten's webpage on the 1st of October 2020. Due to the fact that we downloaded it manually on a specific date, reproduceability might be an issue since COVID cases might be updated. Our task is to data wrangling but also do some statistic analysis and plotting.

```
## $`Antal per dag region`
## # A tibble: 239 x 23
```

```
##    Statistikdatum      Totalt_~1 Bleki~2 Dalarna Gotland Gävle~3 Halland Jämtl~
##    <dttm>                  <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl
##  1 2020-02-04 00:00:00         1       0       0       0       0       0
##  2 2020-02-05 00:00:00         0       0       0       0       0       0
##  3 2020-02-06 00:00:00         0       0       0       0       0       0
##  4 2020-02-07 00:00:00         0       0       0       0       0       0
##  5 2020-02-08 00:00:00         0       0       0       0       0       0
##  6 2020-02-09 00:00:00         0       0       0       0       0       0
##  7 2020-02-10 00:00:00         0       0       0       0       0       0
##  8 2020-02-11 00:00:00         0       0       0       0       0       0
##  9 2020-02-12 00:00:00         0       0       0       0       0       0
## 10 2020-02-13 00:00:00         0       0       0       0       0       0
## # ... with 229 more rows, 15 more variables: Jönköping <dbl>, Kalmar <dbl>,
## #   Kronoberg <dbl>, Norrbotten <dbl>, Skåne <dbl>, Stockholm <dbl>,
## #   Sörmland <dbl>, Uppsala <dbl>, Värmland <dbl>, Västerbotten <dbl>,
## #   Västernorrland <dbl>, Västmanland <dbl>, Västra_Götaland <dbl>,
## #   Örebro <dbl>, Östergötland <dbl>, and abbreviated variable names
## #   1: Totalt_antal_fall, 2: Blekinge, 3: Gävleborg, 4: Jämtland_Härjedalen
##
## $`Antal avlidna per dag`
## # A tibble: 204 x 2
##    Datum_avliden Antal_avlidna
##    <chr>                 <dbl>
##  1 43901                     1
##  2 43902                     0
##  3 43903                     1
##  4 43904                     1
##  5 43905                     2
##  6 43906                     2
##  7 43907                     1
##  8 43908                     6
##  9 43909                     7
## 10 43910                     9
## # ... with 194 more rows
##
## $`Antal intensivvårdade per dag`
## # A tibble: 208 x 2
##    Datum_vårdstart     Antal_intensivvårdade
##    <dttm>                              <dbl>
##  1 2020-03-06 00:00:00                     1
##  2 2020-03-07 00:00:00                     1
##  3 2020-03-08 00:00:00                     1
##  4 2020-03-09 00:00:00                     0
##  5 2020-03-10 00:00:00                     2
##  6 2020-03-11 00:00:00                     1
##  7 2020-03-12 00:00:00                     0
##  8 2020-03-13 00:00:00                     2
##  9 2020-03-14 00:00:00                     6
## 10 2020-03-15 00:00:00                     5
## # ... with 198 more rows
##
## $`Totalt antal per region`
## # A tibble: 21 x 5
```

```
##    Region              Totalt_antal_fall Fall_per_100000_inv Totalt_an~1 Total~
##    <chr>                           <dbl>               <dbl>       <dbl>   <dbl
##  1 Blekinge                          712                446.           9       1
##  2 Dalarna                          2543                883.          67      17
##  3 Gotland                           330                553.           7
##  4 Gävleborg                        3379               1176.          76      16
##  5 Halland                          2576                772.          41       8
##  6 Jämtland Härjedalen              1289                985.          20       6
##  7 Jönköping                        5375               1478.          96      18
##  8 Kalmar                            929                378.          31       6
##  9 Kronoberg                        1705                846.          25      12
## 10 Norrbotten                       1750                700.          60       8
## # ... with 11 more rows, and abbreviated variable names
## #   1: Totalt_antal_intensivvårdade, 2: Totalt_antal_avlidna
##
## $`Totalt antal per kön`
## # A tibble: 3 x 4
##   Kön            Totalt_antal_fall Totalt_antal_intensivvårdade Totalt_antal_a~
##   <chr>                      <dbl>                        <dbl>            <dbl
## 1 Man                        40380                         1897              322
## 2 Kvinna                     52476                          708              267
## 3 Uppgift saknas                 7                            0
## # ... with abbreviated variable name 1: Totalt_antal_avlidna
##
## $`Totalt antal per åldersgrupp`
## # A tibble: 11 x 4
##    Åldersgrupp    Totalt_antal_fall Totalt_antal_intensivvårdade Totalt_antal_~
##    <chr>                      <dbl>                        <dbl>           <dbl
##  1 Ålder_0_9                    710                            8
##  2 Ålder_10_19                 4783                           18
##  3 Ålder_20_29                15700                           97               1
##  4 Ålder_30_39                14469                          119               1
##  5 Ålder_40_49                15143                          289               4
##  6 Ålder_50_59                16129                          661              16
##  7 Ålder_60_69                 9166                          781              40
##  8 Ålder_70_79                 6259                          515             126
##  9 Ålder_80_89                 6822                          113             244
## 10 Ålder_90_plus               3661                            4             153
## 11 Uppgift saknas                21                            0
## # ... with abbreviated variable name 1: Totalt_antal_avlidna
##
## $`Veckodata Region`
## # A tibble: 693 x 10
##    veckonummer Region   Antal_~1 Kum_a~2 Antal~3 Kum_a~4 Antal~5 Kum_a~6 Antal~
##          <dbl> <chr>       <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl
##  1           6 Blekinge        0       0       0       0       0       0
##  2           7 Blekinge        0       0       0       0       0       0
##  3           8 Blekinge        0       0       0       0       0       0
##  4           9 Blekinge        0       0       0       0       0       0
##  5          10 Blekinge        0       0       0       0       0       0
##  6          11 Blekinge       10      10       0       0       0       0
##  7          12 Blekinge        2      12       0       0       0       0
##  8          13 Blekinge        9      21       1       1       1       1
```

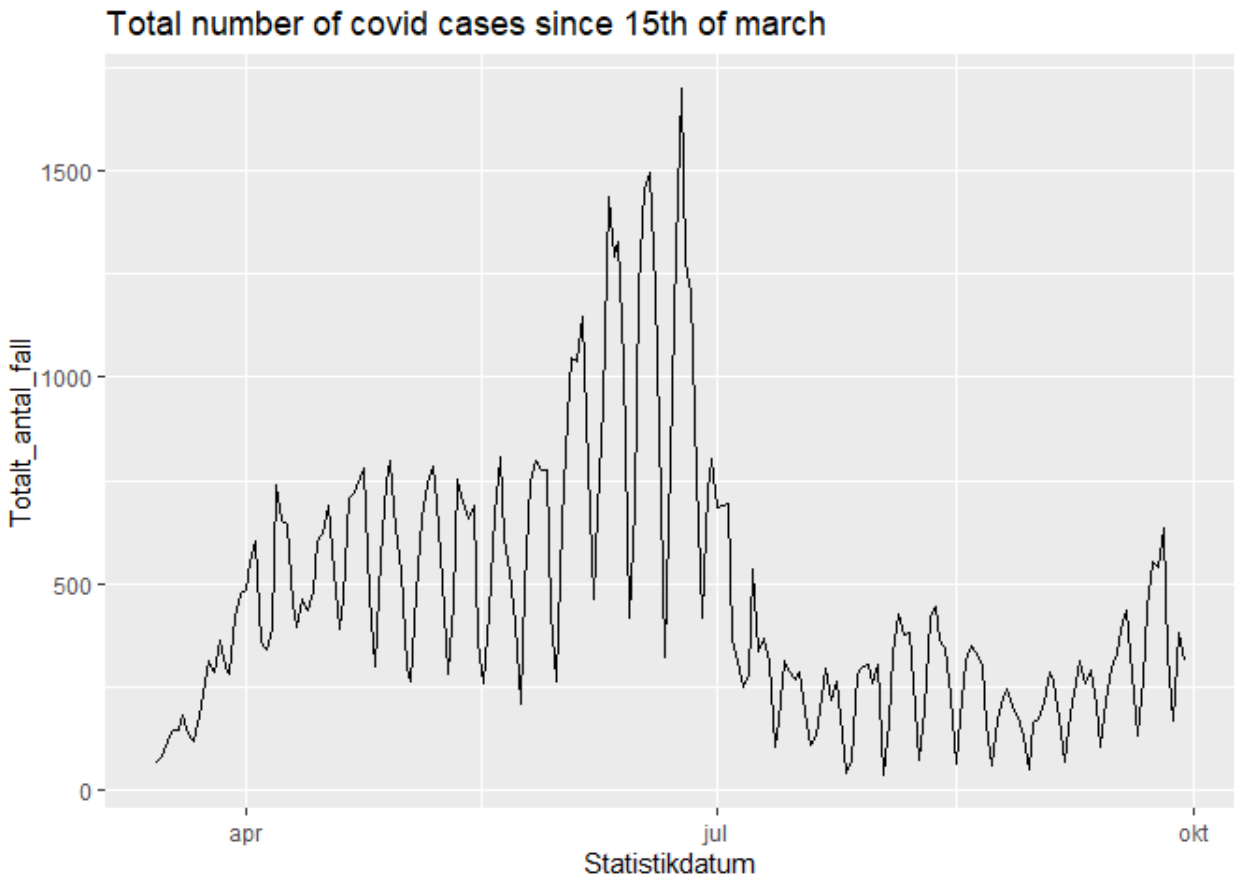
```
##  9          14 Blekinge       15      36       1       2       0       1
## 10          15 Blekinge        6      42       0       2       1       2
## # ... with 683 more rows, 1 more variable: Kum_fall_100000inv <dbl>, and
## #   abbreviated variable names 1: Antal_fall_vecka, 2: Kum_antal_fall,
## #   3: Antal_intensivvårdade_vecka, 4: Kum_antal_intensivvårdade,
## #   5: Antal_avlidna_vecka, 6: Kum_antal_avlidna, 7: Antal_fall_100000inv_vecka
##
## $`Veckodata Kommun_stadsdel`
## # A tibble: 10,626 x 9
##    veckonummer KnKod KnNamn Stadsdel Kommun_st~1 tot_a~2 antal~3 tot_a~4 nya_f~
##          <dbl> <chr> <chr>  <lgl>    <chr>         <dbl>   <dbl> <chr>   <chr>
##  1           6 1440  Ale    NA       Ale               0       0 0       0
##  2           7 1440  Ale    NA       Ale               0       0 0       0
##  3           8 1440  Ale    NA       Ale               0       0 0       0
##  4           9 1440  Ale    NA       Ale               0       0 0       0
##  5          10 1440  Ale    NA       Ale               0       0 0       0
##  6          11 1440  Ale    NA       Ale              NA      NA <15     <15
##  7          12 1440  Ale    NA       Ale              NA      NA <15     <15
##  8          13 1440  Ale    NA       Ale              NA      NA <15     <15
##  9          14 1440  Ale    NA       Ale               6       3 19      9
## 10          15 1440  Ale    NA       Ale               9       3 27      8
## # ... with 10,616 more rows, and abbreviated variable names 1: Kommun_stadsdel,
## #   2: tot_antal_fall_per10000inv, 3: antal_fall_per10000_inv,
## #   4: tot_antal_fall, 5: nya_fall_vecka
##
## $`FOHM 30 Sep 2020`
## # A tibble: 1 x 1
##   Information
##   <chr>
## 1 Data uppdateras vardagar kl 14.00 med data fram till föregående dag. Veckodat
```

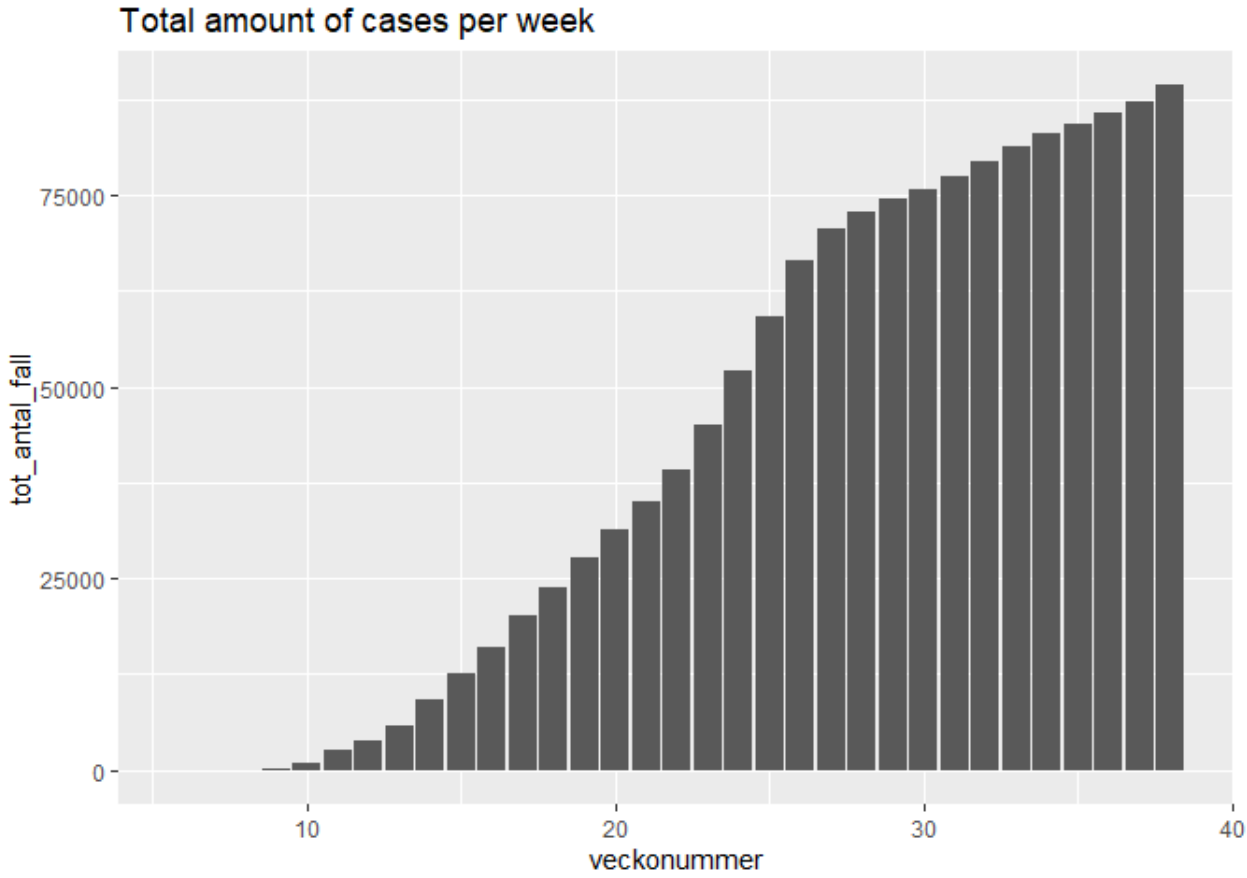
| Datum_avliden | Antal_avlidna |
|---|---|
| 2020-03-11 | 1 |
| 2020-03-12 | 0 |
| 2020-03-13 | 1 |
| 2020-03-14 | 1 |
| 2020-03-15 | 2 |
| 2020-09-25 | 3 |
| 2020-09-26 | 2 |
| 2020-09-27 | 0 |
| 2020-09-28 | 0 |

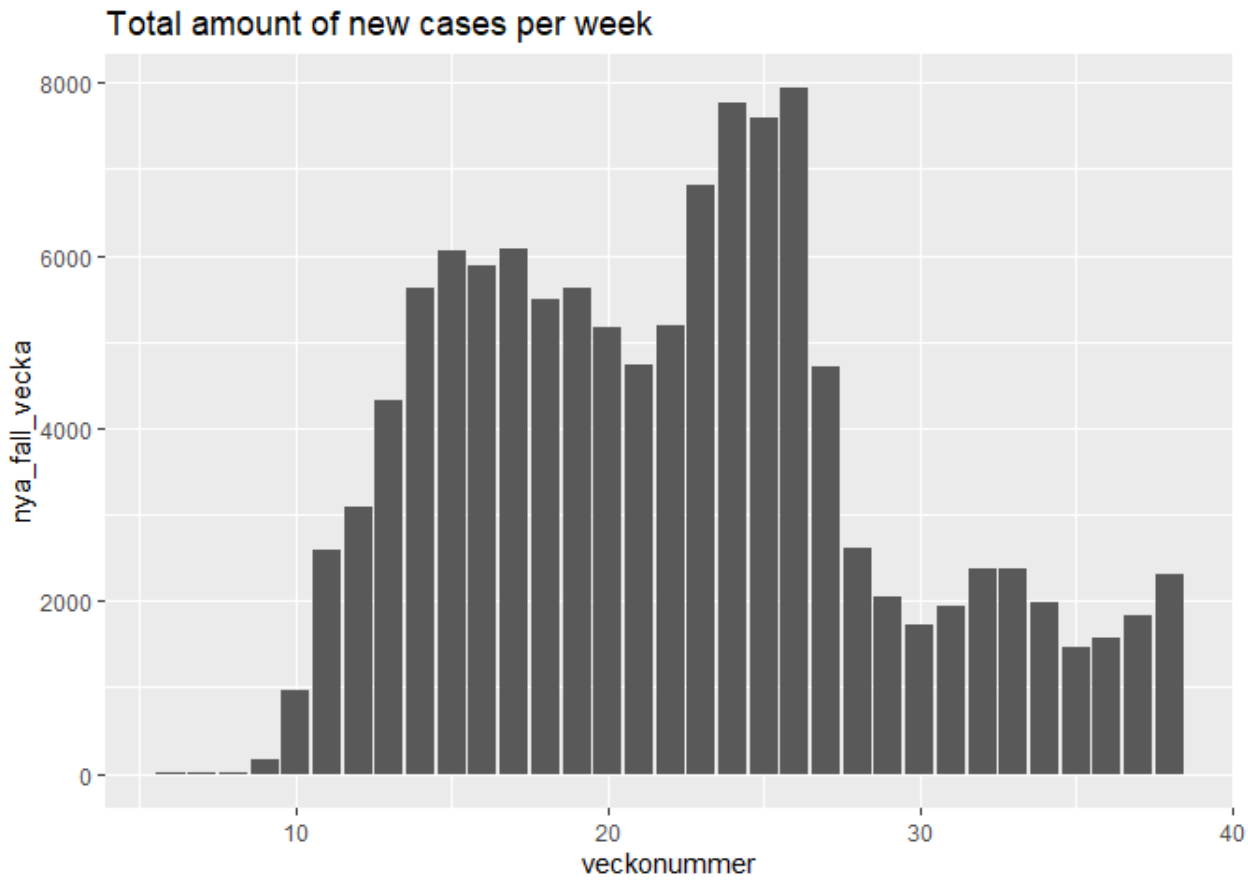
| Datum_avliden | Antal_avlidna |
|---|---|
| 2020-09-29 | 0 |

```
## # A tibble: 10,626 x 9
##    veckonummer KnKod KnNamn Stadsdel Kommun_st~1 tot_a~2 antal~3 tot_a~4 nya_f~
##          <dbl> <chr> <chr>  <lgl>    <chr>         <dbl>   <dbl>   <dbl>   <dbl
##  1           6 1440  Ale    NA       Ale               0       0       0
##  2           7 1440  Ale    NA       Ale               0       0       0
##  3           8 1440  Ale    NA       Ale               0       0       0
##  4           9 1440  Ale    NA       Ale               0       0       0
##  5          10 1440  Ale    NA       Ale               0       0       0
##  6          11 1440  Ale    NA       Ale              NA      NA      15       1
##  7          12 1440  Ale    NA       Ale              NA      NA      15       1
##  8          13 1440  Ale    NA       Ale              NA      NA      15       1
##  9          14 1440  Ale    NA       Ale               6       3      19
## 10          15 1440  Ale    NA       Ale               9       3      27
## # ... with 10,616 more rows, and abbreviated variable names 1: Kommun_stadsdel,
## #   2: tot_antal_fall_per10000inv, 3: antal_fall_per10000_inv,
## #   4: tot_antal_fall, 5: nya_fall_vecka
```

```
##   Totalt_antal_fall            Blekinge             Dalarna             Gotland
##               92863                 712                2543                 330
##           Gävleborg             Halland Jämtland_Härjedalen           Jönköping
##                3379                2576                1289                5375
##              Kalmar           Kronoberg          Norrbotten             Skåne
##                 929                1705                1750                5861
##           Stockholm            Sörmland             Uppsala            Värmland
##               25146                2521                4072                1275
##        Västerbotten      Västernorrland         Västmanland     Västra_Götaland
##                1030                1919                3121               20247
##              Örebro        Östergötland
##                2991                4092
```


Total number of covid cases since 15th of march
Totalt_antal_fall
1500
1000
500
0
apr
jul
okt
Statistikdatum


Total amount of cases per week
tot_antal_fall
75000
50000
25000
0
10
20
30
40
veckonummer




We see that there are 9 sheets in total. Each sheet tells us something different, but to make a summarize of all the sheets can we say that the data set shows us the amount of covid cases/deaths per day. It also divides the amount of cases in genders, regions and age. To end the summarize we also get the date on which the cases occurred also the amount of people who got intense caring for every day. Observe that this data set is only limited and does not include newest cases.

We do not seem to get any problems when we try to display the first and last 5 rows. What i did was to first select the 2nd sheet, “antal avlidna per dap” and then revomed the last row because it didnt contain a date but it did contain a number. After removing that row I just removed the rows given as a vector from 6:198 and what’s left is the 5 first and last rows.

The thing read_excel does is to guess the column types based on what the columns are in excel. In our case though that column does not have any values except from a title. So what it does instead is to interpreter as logical and give us “NA”. To fix this problem we can simply remove the column by skiping the col_types.

The reason why we get chr instead of dbl (a floating point number) is because of the sign “<”. We get this sign in both tot_antal_fall and nya_fall_vecka. One thing that can be done in order to fix this is to maybe approximate the cases and give them the limit of 15 in this we can remove the “<” sign and in this way we can then change chr to a dbl. That’s the option i choose to go for.

The total amount of cases can be computed adding the code summarize the sum och the column “totalt_antal_fall” and we get that in total there was 92863 cases registered. By arranging the table we can say that stockholm had the most cases of covid and gotland got the lowest. One thing we cant determine by the table is how much percentage of the cases represents each region. This could be misleading in a way because one wont know if it is a huge number or not for that certain region. One thing one can do is find out the population of that region and use that information to find out the actual percentage based on the cases.

The dataframe that produced the figure in c is suppose to have 10626 rows and sorted from week 6 to week 38. What ggplot did without telling us is to interpreter each week as a variable from 6-38. Else it would was to interpreter each week number as an x-variable that means we will have had over 10000 variables on the x-axis (we did avoid this by changing chr to dbl). This would had give our table c a very tight and clumsy character.

```
## R version 4.1.2 (2021-11-01)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19045)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=Swedish_Sweden.1252  LC_CTYPE=Swedish_Sweden.1252
## [3] LC_MONETARY=Swedish_Sweden.1252 LC_NUMERIC=C
## [5] LC_TIME=Swedish_Sweden.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] readxl_1.4.1    forcats_0.5.2   stringr_1.4.1   dplyr_1.0.10
##  [5] purrr_0.3.4     readr_2.1.3     tidyr_1.2.1     tibble_3.1.8
##  [9] ggplot2_3.4.0   tidyverse_1.3.2
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_1.2.0    xfun_0.29           haven_2.5.1
##  [4] gargle_1.2.1        colorspace_2.0-3    vctrs_0.5.0
##  [7] generics_0.1.3      htmltools_0.5.3     yaml_2.3.6
## [10] utf8_1.2.2          rlang_1.0.6         pillar_1.8.1
## [13] withr_2.5.0         glue_1.6.2          DBI_1.1.3
## [16] dbplyr_2.2.1        modelr_0.1.9        lifecycle_1.0.3
## [19] munsell_0.5.0       gtable_0.3.1        cellranger_1.1.0
## [22] rvest_1.0.3         evaluate_0.18       labeling_0.4.2
## [25] knitr_1.40          tzdb_0.3.0          fastmap_1.1.0
## [28] fansi_1.0.3         highr_0.9           broom_1.0.1
## [31] scales_1.2.1        backports_1.4.1     googlesheets4_1.0.1
## [34] jsonlite_1.8.3      farver_2.1.1        fs_1.5.2
## [37] hms_1.1.2           digest_0.6.29       stringi_1.7.6
```

```
## [40] grid_4.1.2          cli_3.4.1           tools_4.1.2
## [43] magrittr_2.0.3      crayon_1.5.2        pkgconfig_2.0.3
## [46] ellipsis_0.3.2      xml2_1.3.3          reprex_2.0.2
## [49] googledrive_2.0.0   lubridate_1.9.0     timechange_0.1.1
## [52] assertthat_0.2.1    rmarkdown_2.18      httr_1.4.4
## [55] rstudioapi_0.14     R6_2.5.1            compiler_4.1.2
```